**Major Project**

1. Executive Summary

1.1 Project Overview

To develop the System to analyze the log data (In XML format) of government progress of various development activities.

1.2 Purpose and Scope of this Specification

The purpose of this project is to capture the data for analyzing the progress of various activities.

In scope The following requirement will be addressed in phase 1 of Project:

• Developing system to handle the incoming log feed and store the information in Hadoop Cluster (Flume)

• Analyze the data and understand the progress

• Store the results in Hbase/RDBMS Out of scope We can use this data and visualization and get more insights

2. Product/Service Description

2.1 Assumptions Log will be generated in XML format and stored in a server

2.2 Constraints

Describe any item that will constrain the design options, including

• This system may not be used for searching for now. But it will be used for analysis and saving the relevant information as of now

• System will be using Hbase as a database Big Data and Hadoop Development

3. Requirements

• The FLUME job which will format the data and place the data to HDFS

• Pig/MapReduce job for parsing the XML data.

• Create Pig scripts/MapReduce jobs to analyze the data

• Create the Sqoop job to store the data in database

Priority Definitions The following definitions are intended as a guideline to prioritize requirements.

• Priority 1 – Create FLUME job for fetching log files from spool directory the data

• Priority 2 – MapReduce/pig job to preprocess Download the dataset using the below link:

Link: <https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing>

Refer the below steps to understand the actual steps to create the above project.

Step 1: Copy dataset from local file system to HDFS using flume.

Note: use the conf file by downloading from below link. Click here to download Command: flume-agent agent –n agent1 –c conf –f Big Data and Hadoop Development

Step 2: Input file is in the XML format use Map reduce or pig to parse the data and get the results for the below problem statements.

4. Problem statement

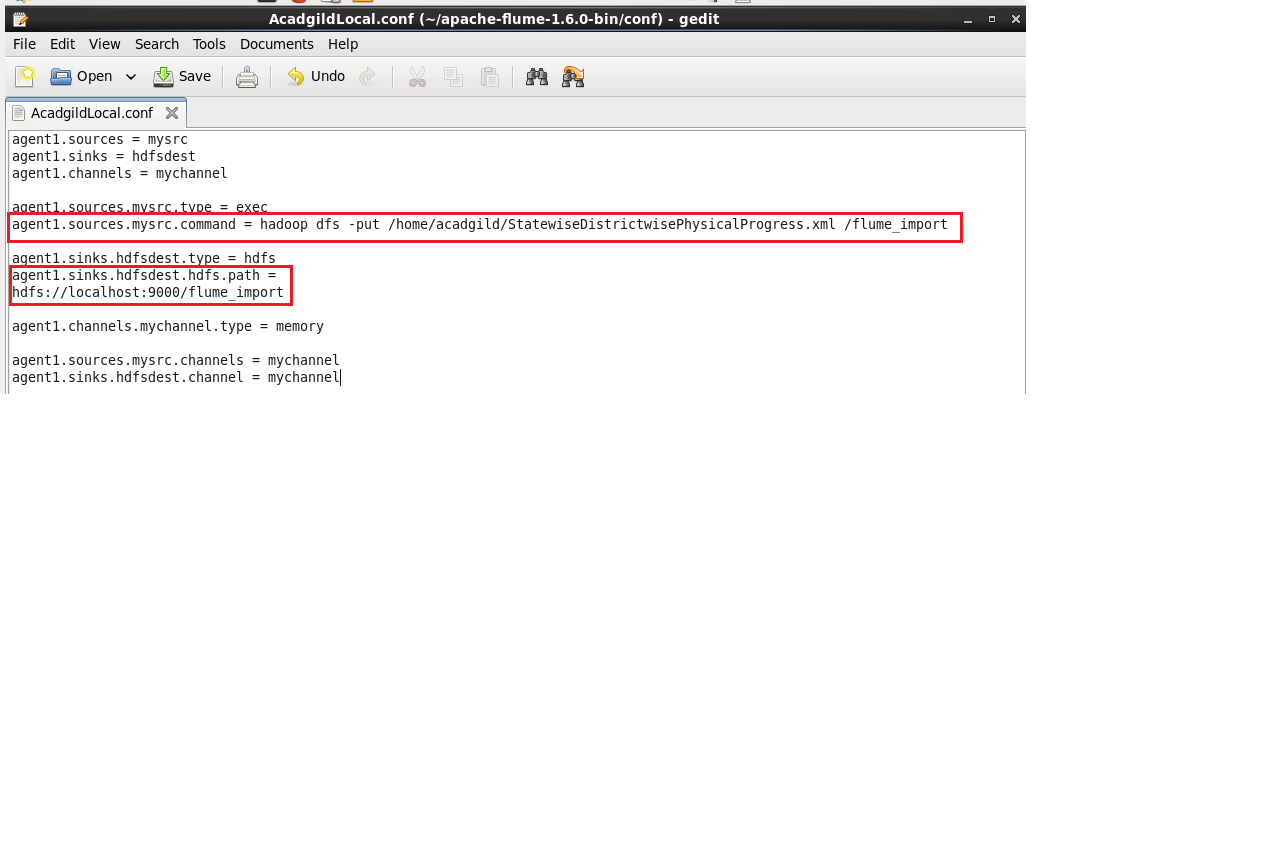
1. Find out the districts who achieved 100 percent objective in BPL cards Export the results to mysql using sqoop

2. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards. Export the results to MySQL using Sqoop.

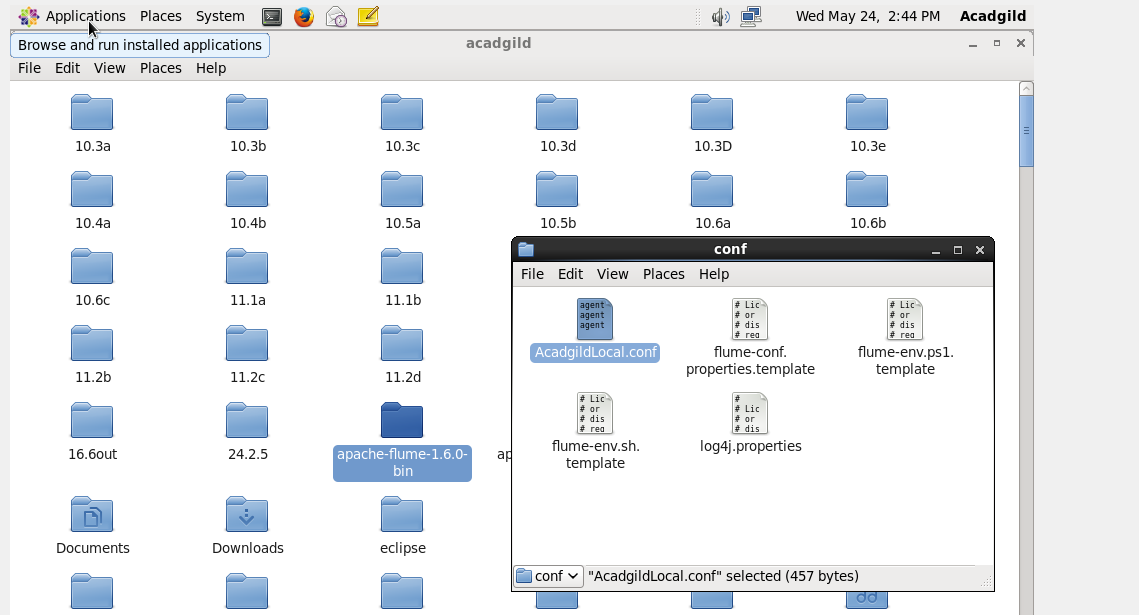
**Solution**

**STEP 1:Copy dataset from local file system to HDFS using Flume**

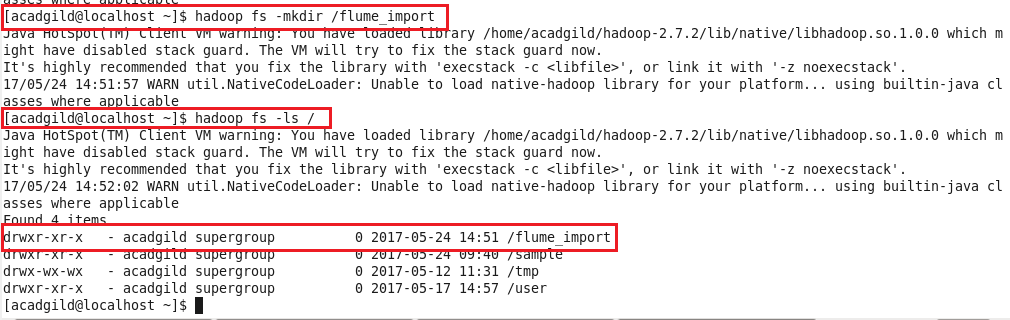
Conf file AcadgildLocal.conf is downloaded from the given link.



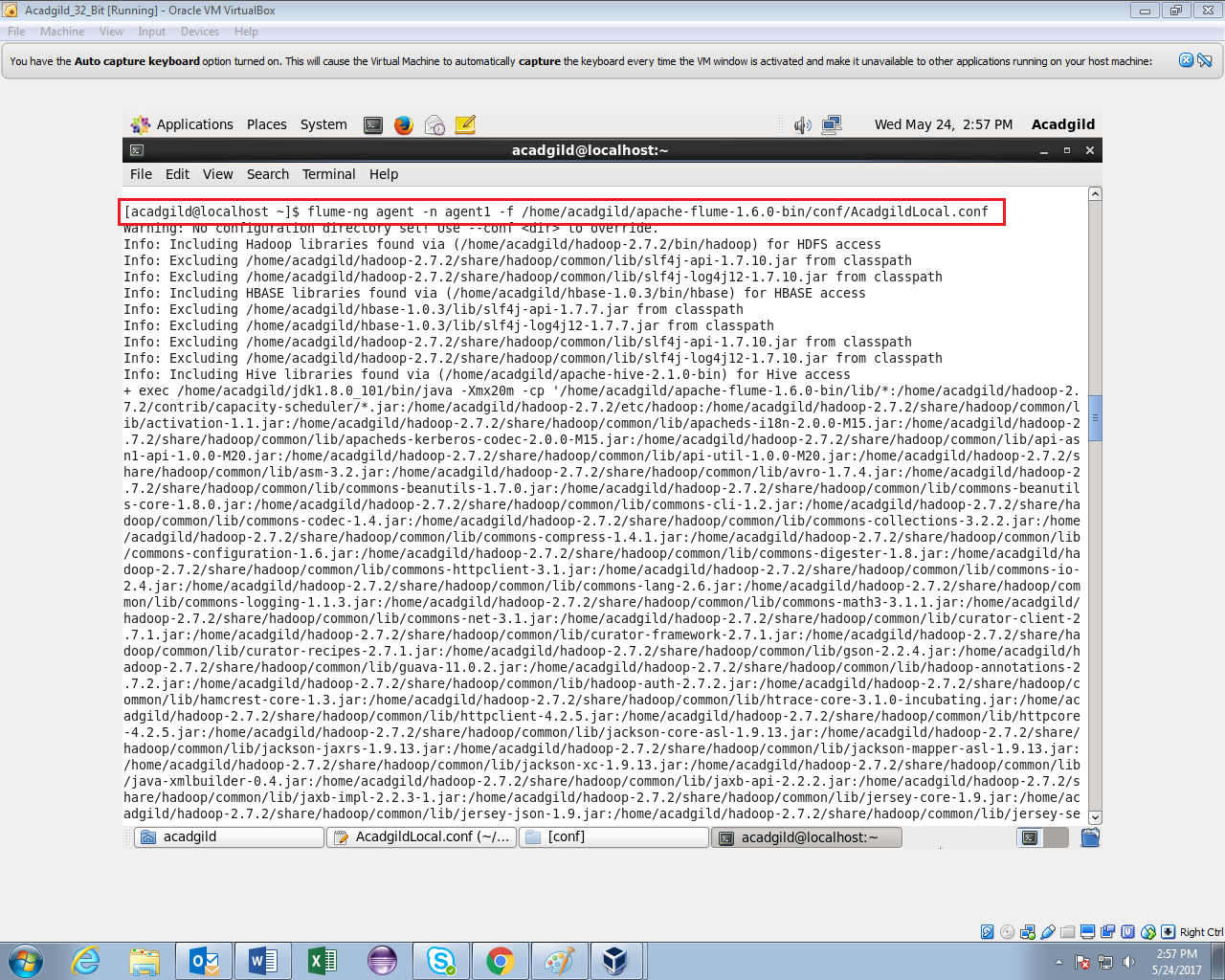
The AcadgildLocal.conf file is moved to apache flume folder.



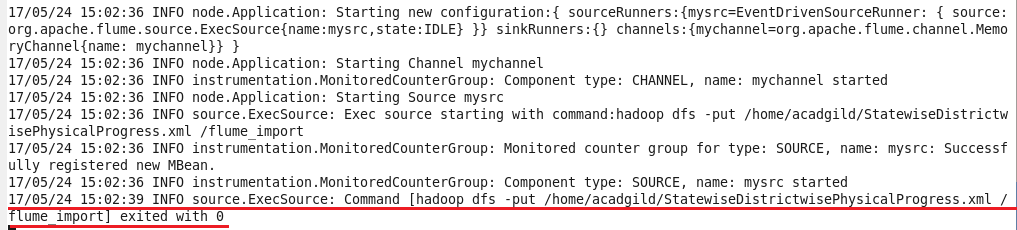
A directory with name flume\_import is created in HDFS as it is mentioned in the conf file.



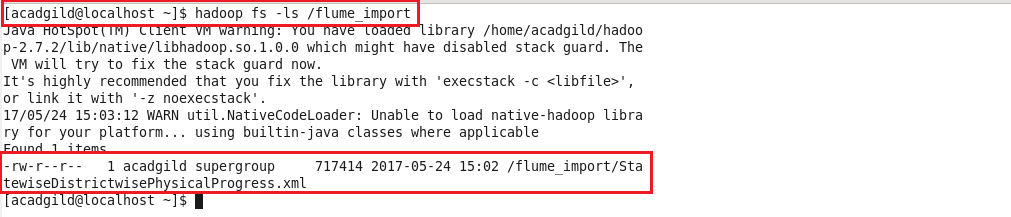
The following command is used to import data into HDFS using flume agent.



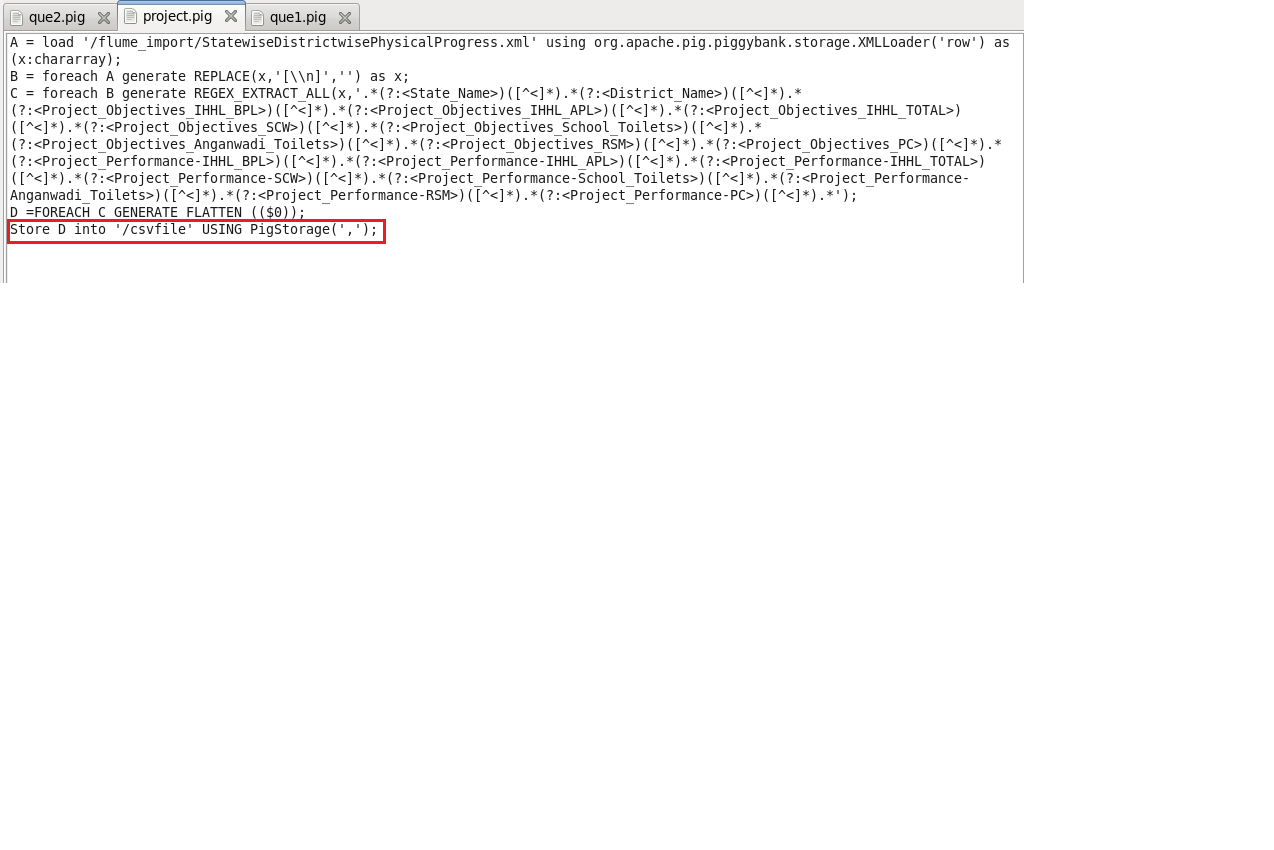
In the screenshot it is seen that flume\_import exited with 0 which means dataset is imported in hdfs.



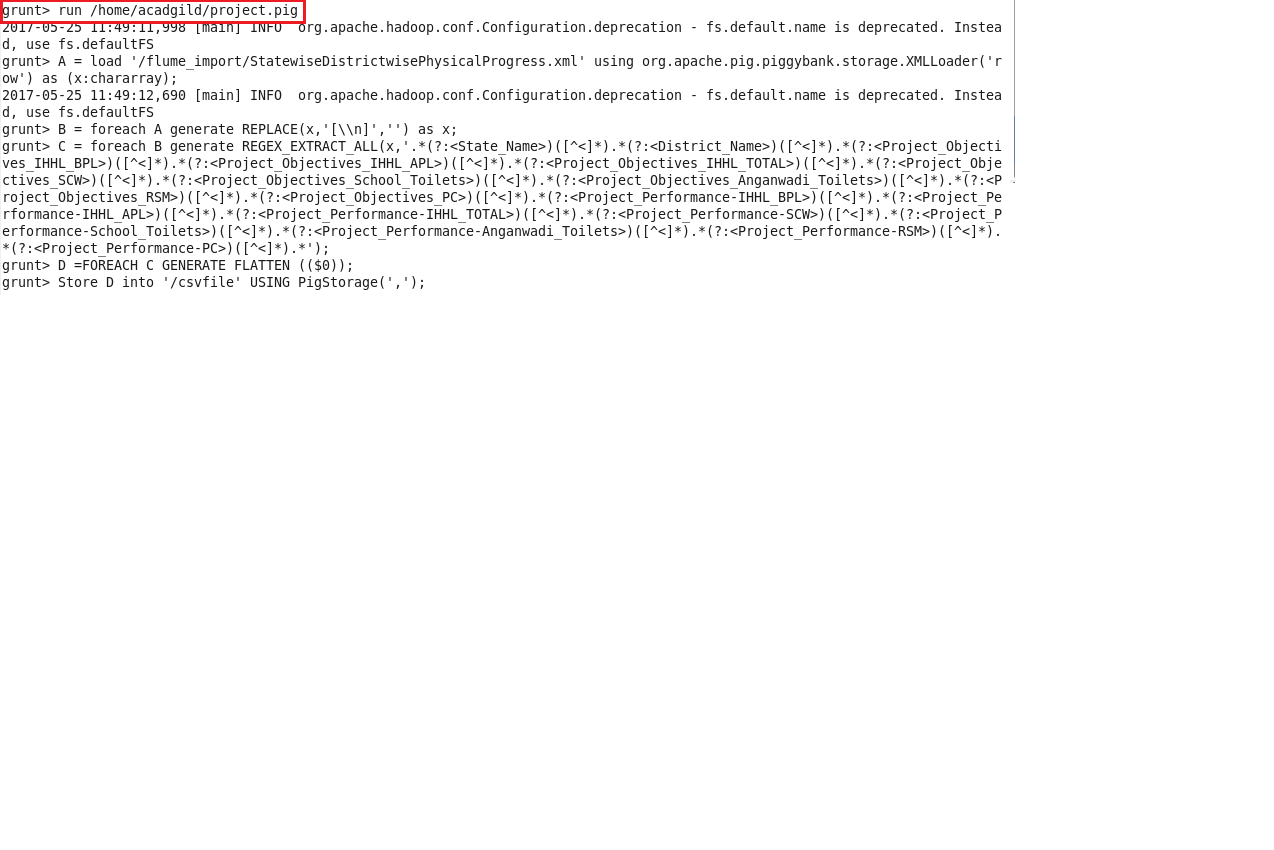
The flume\_import directory is created with the dataset file in xml format.



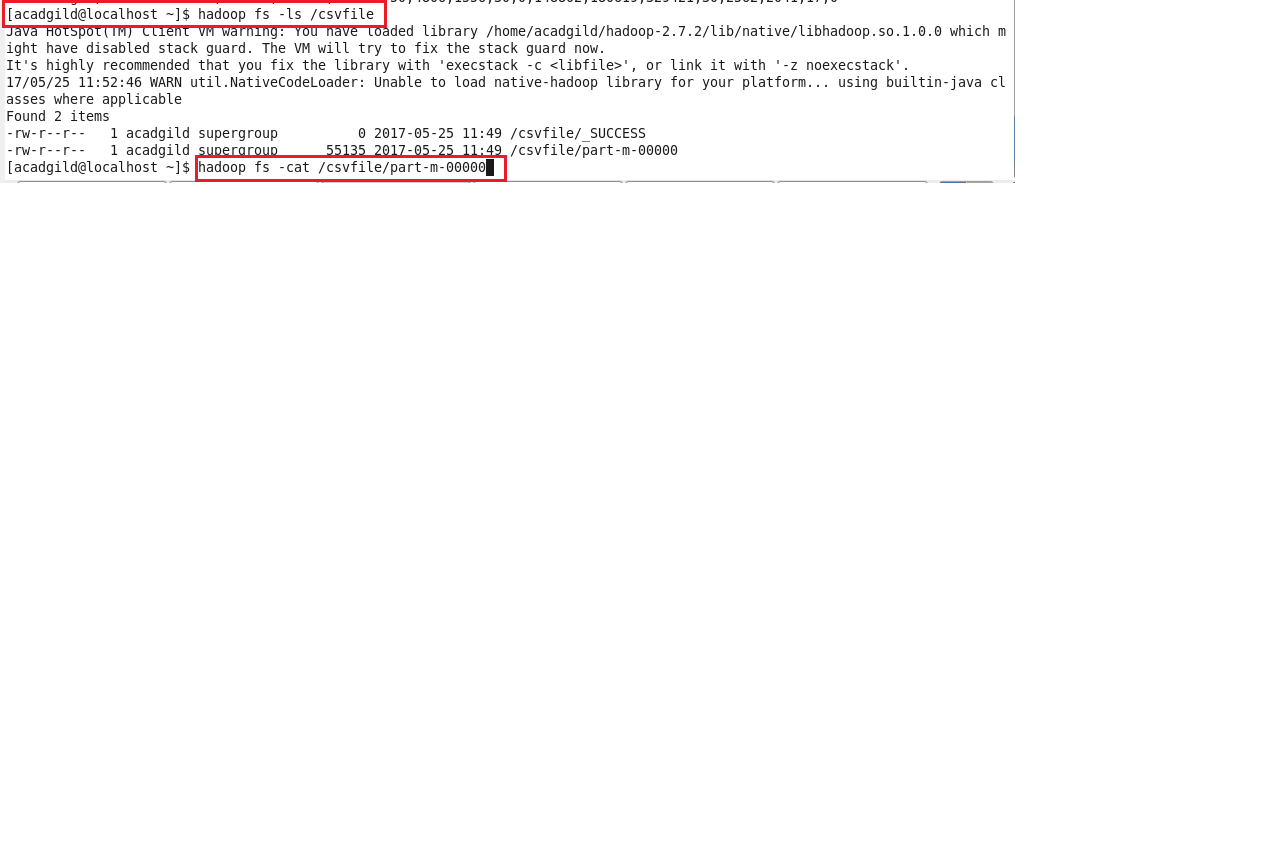
**Step 2: As the input file is in XML format,Pig script is used to parse the data into csv file.The parsed data in csv format is stored in csvfile directory.**



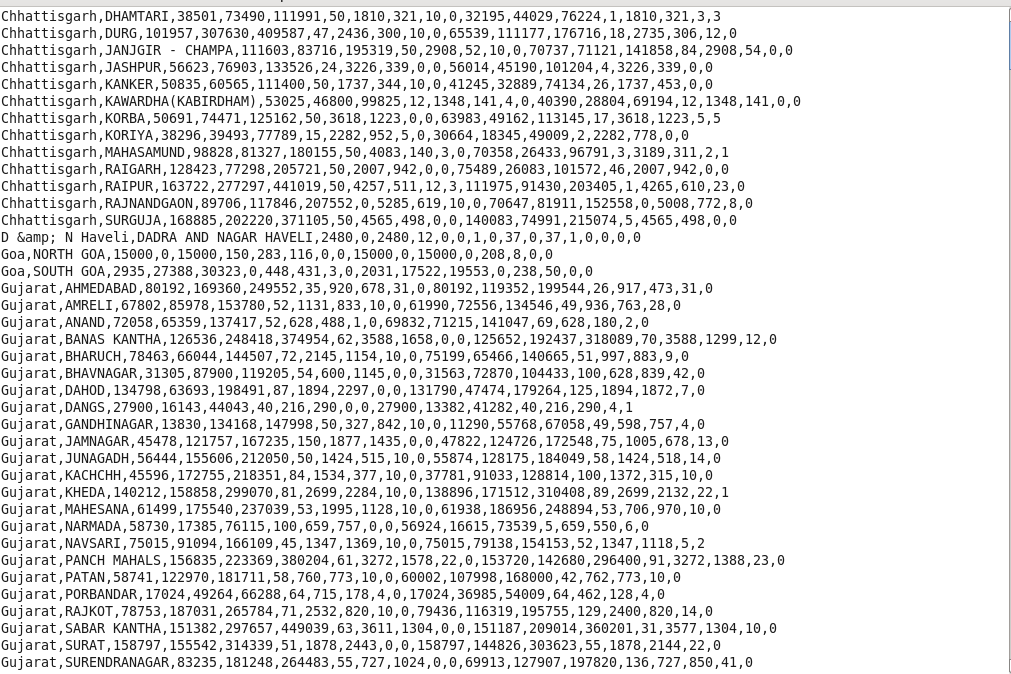
**Running Pig Script**

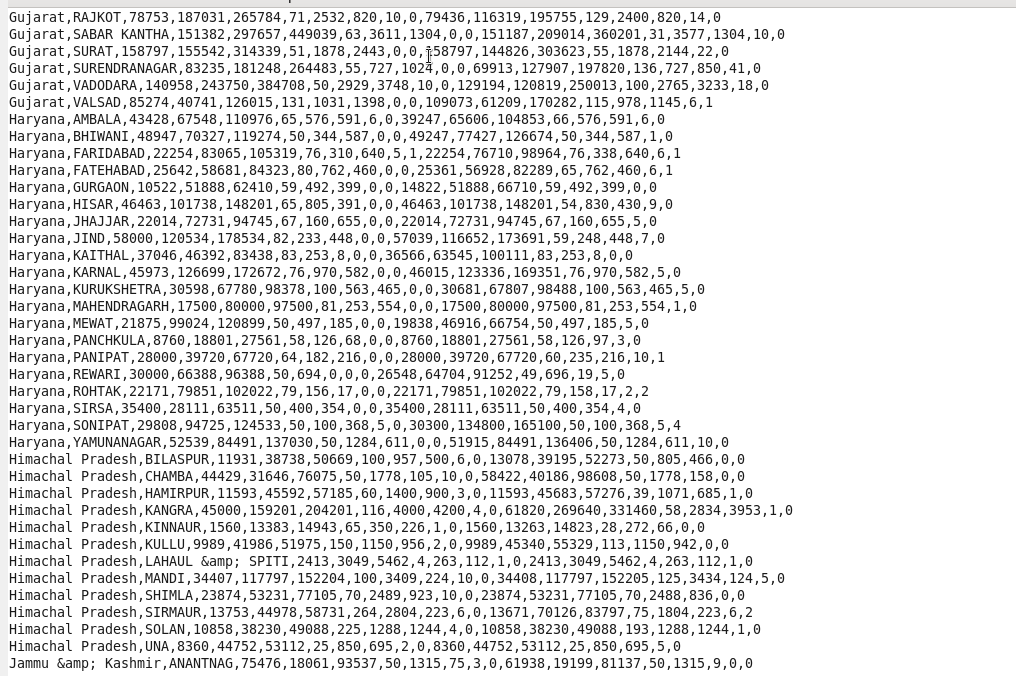


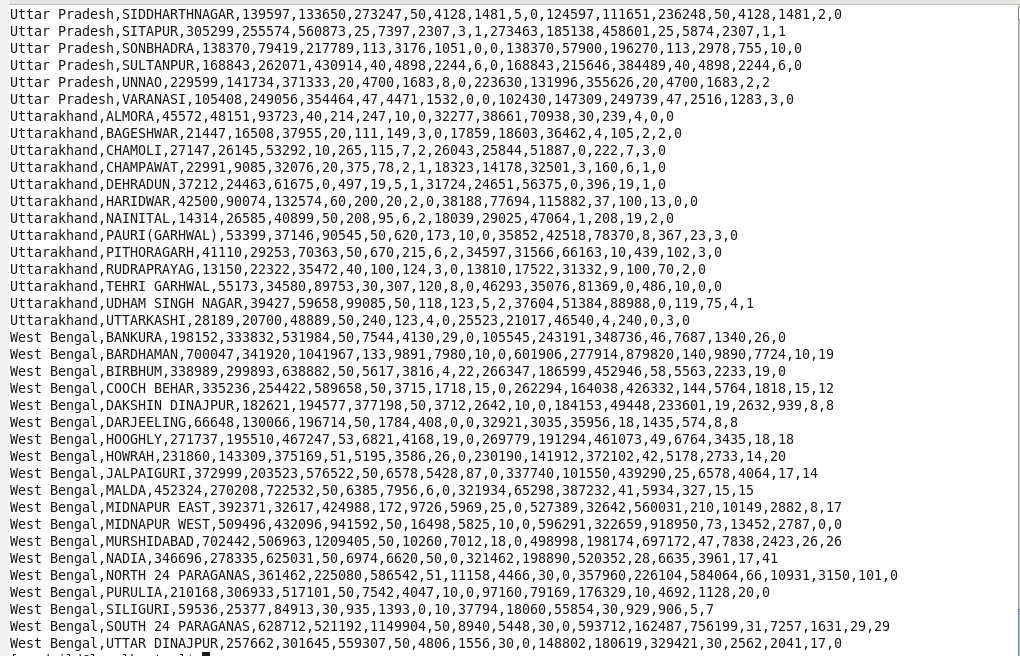
The directory csvfile is created with input file in csv format.



Using cat command the CSV input file is viewed from directory.





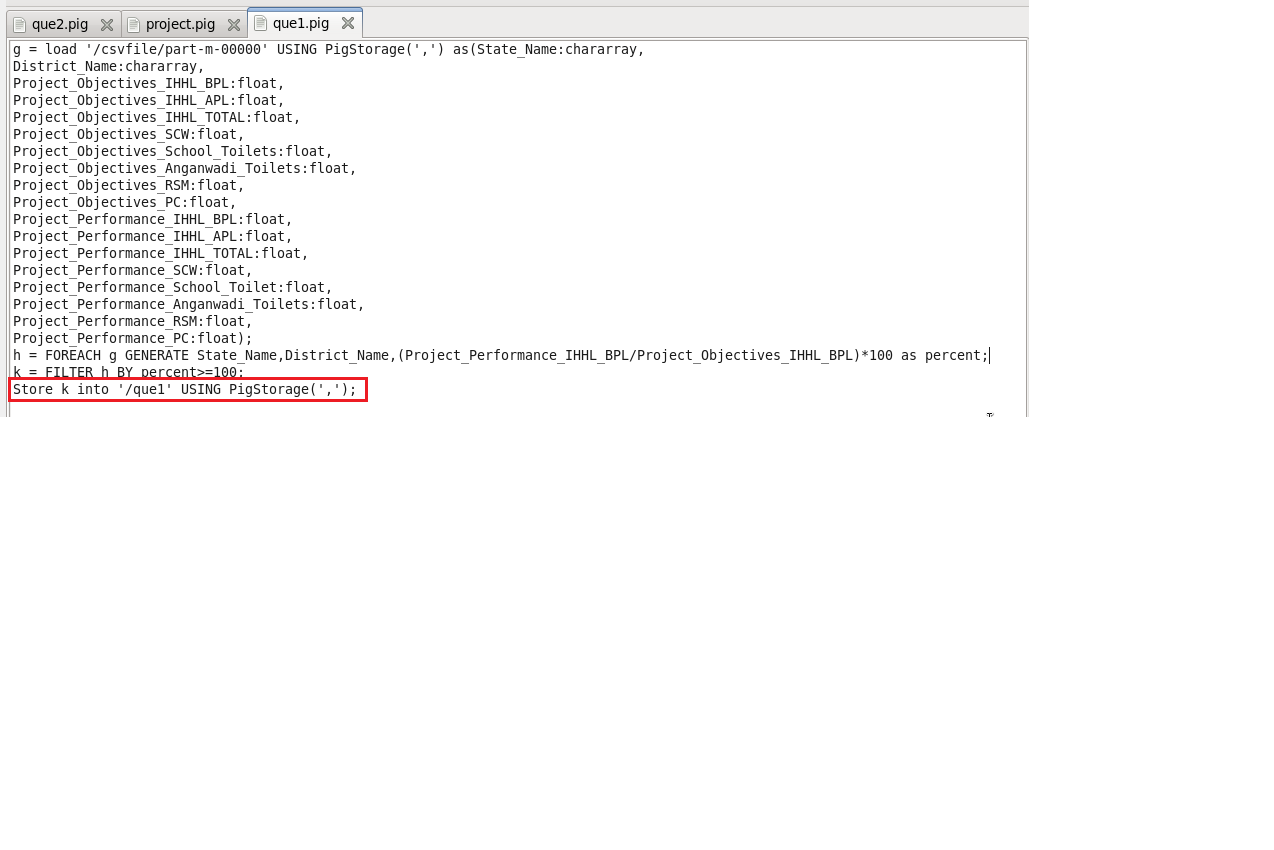


**Problem statement**

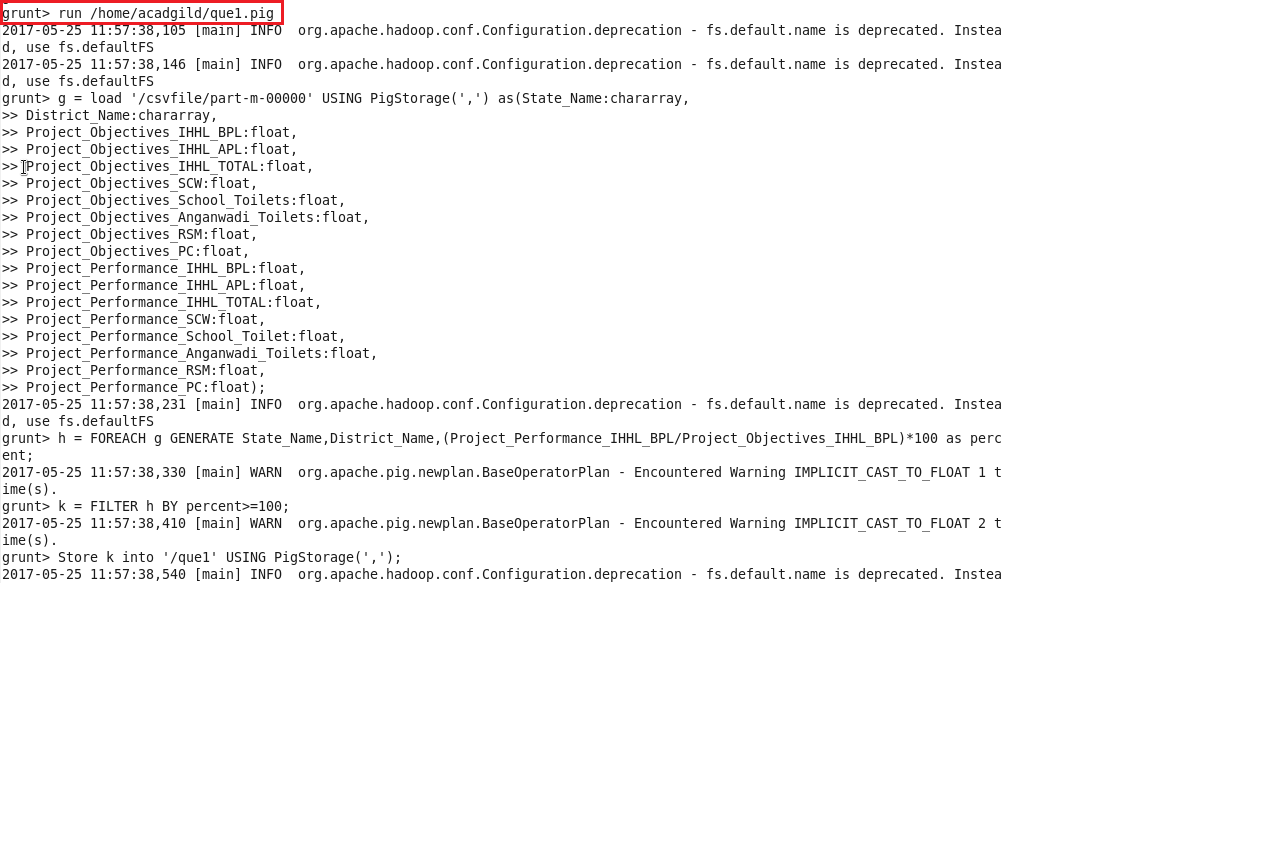
**1. Find out the districts who achieved 100 percent objective in BPL cards Export the results to mysql using sqoop**

**Pig script**

In this,we have loaded the dataset,calculated the percentage and filtered the records having percentage greater than 100.The required columns are generated.The output is stored in que1 directory.



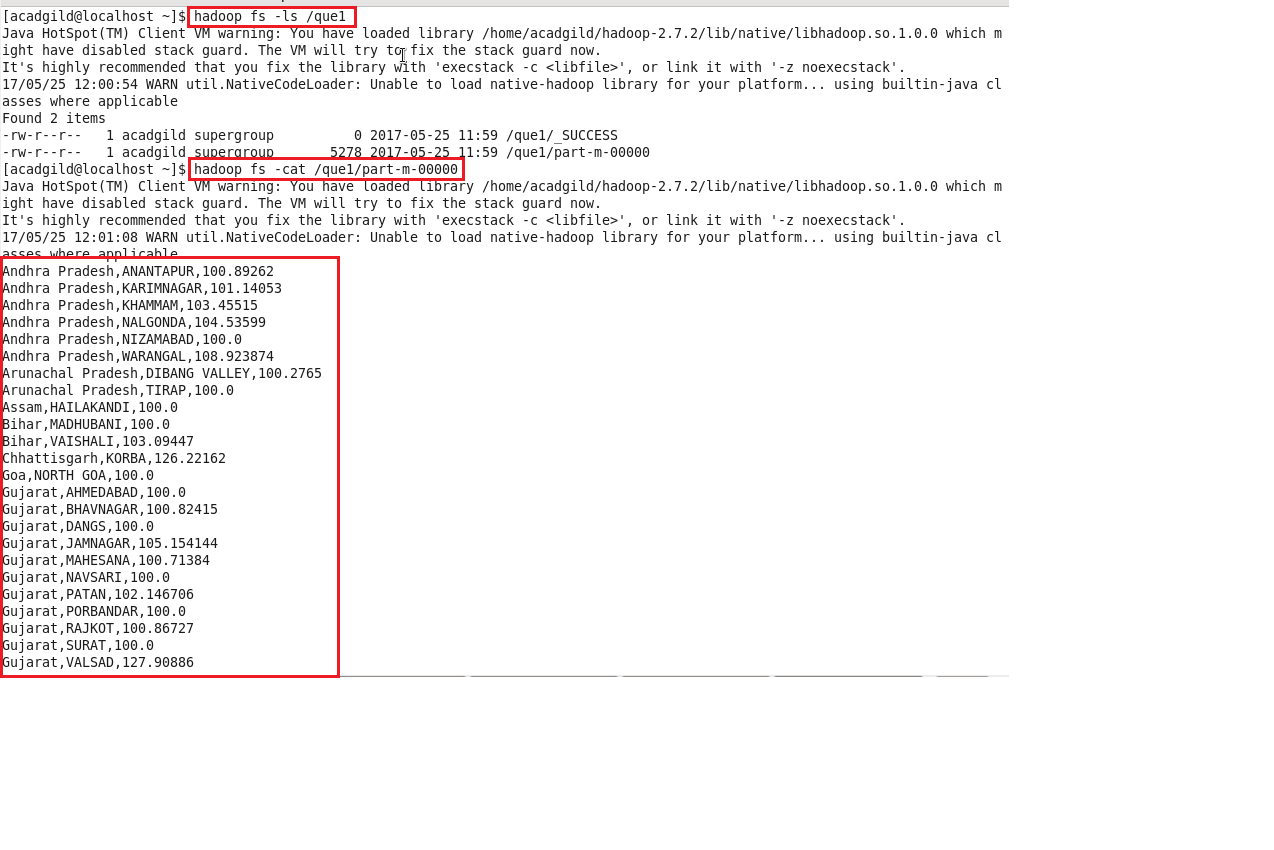
**Running Pig script**

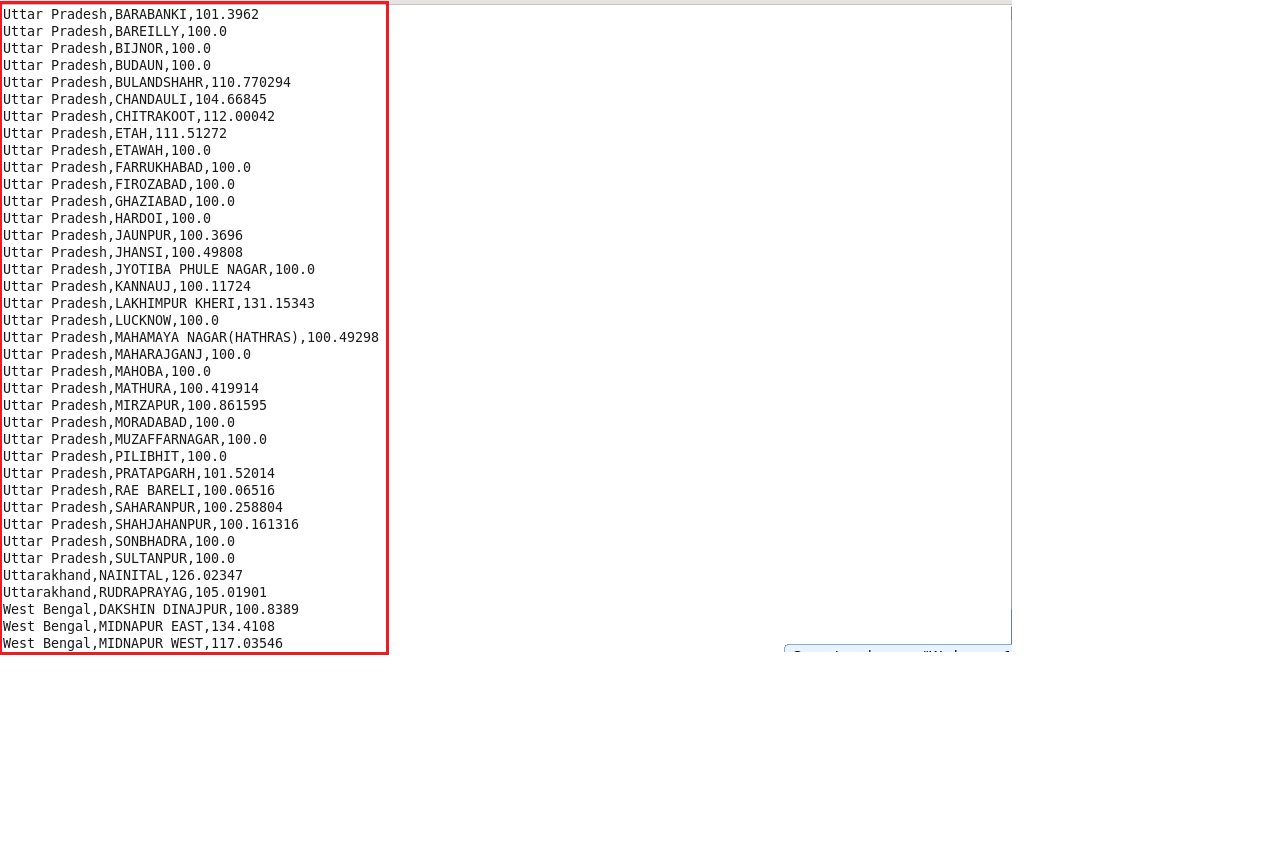


Pig script is executed successfully.



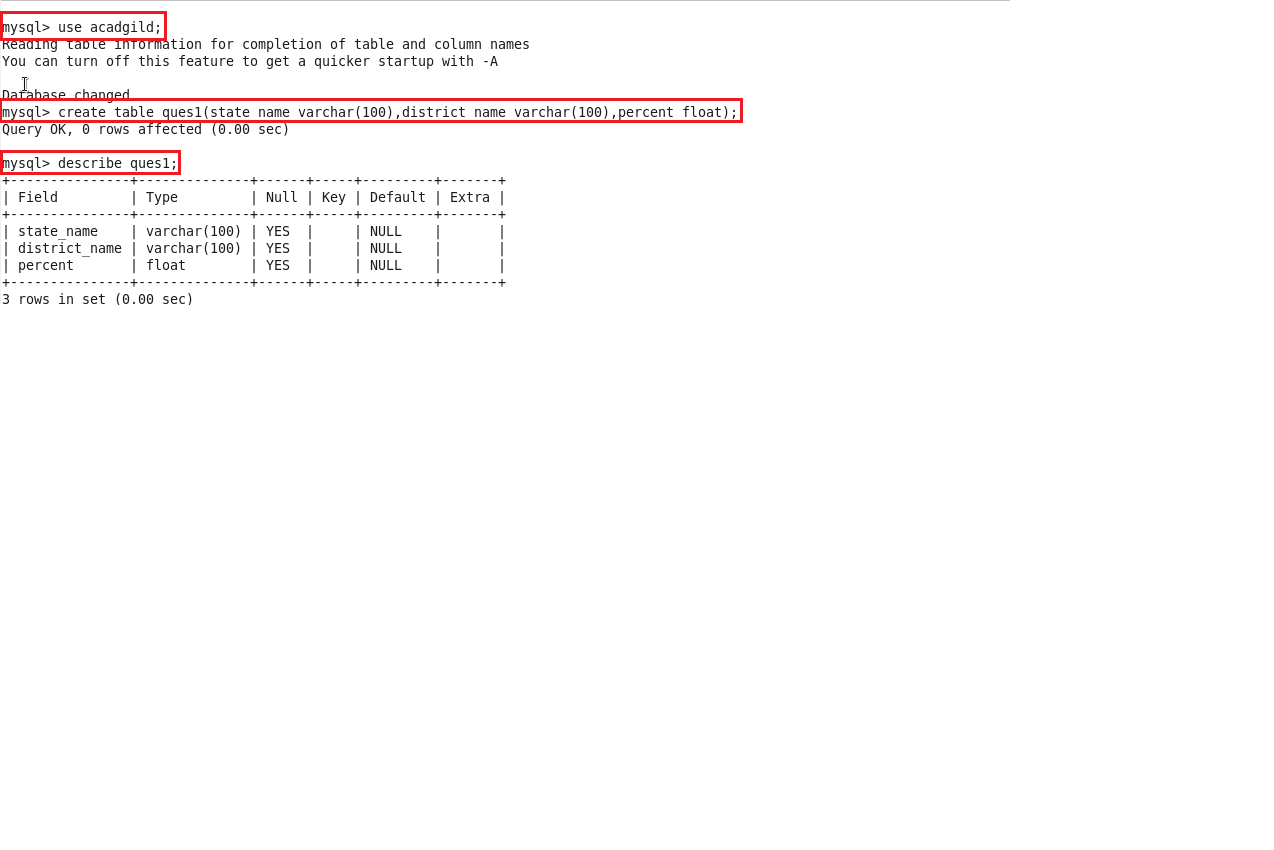
The output records are stored in que1 directory and viewed using cat command.



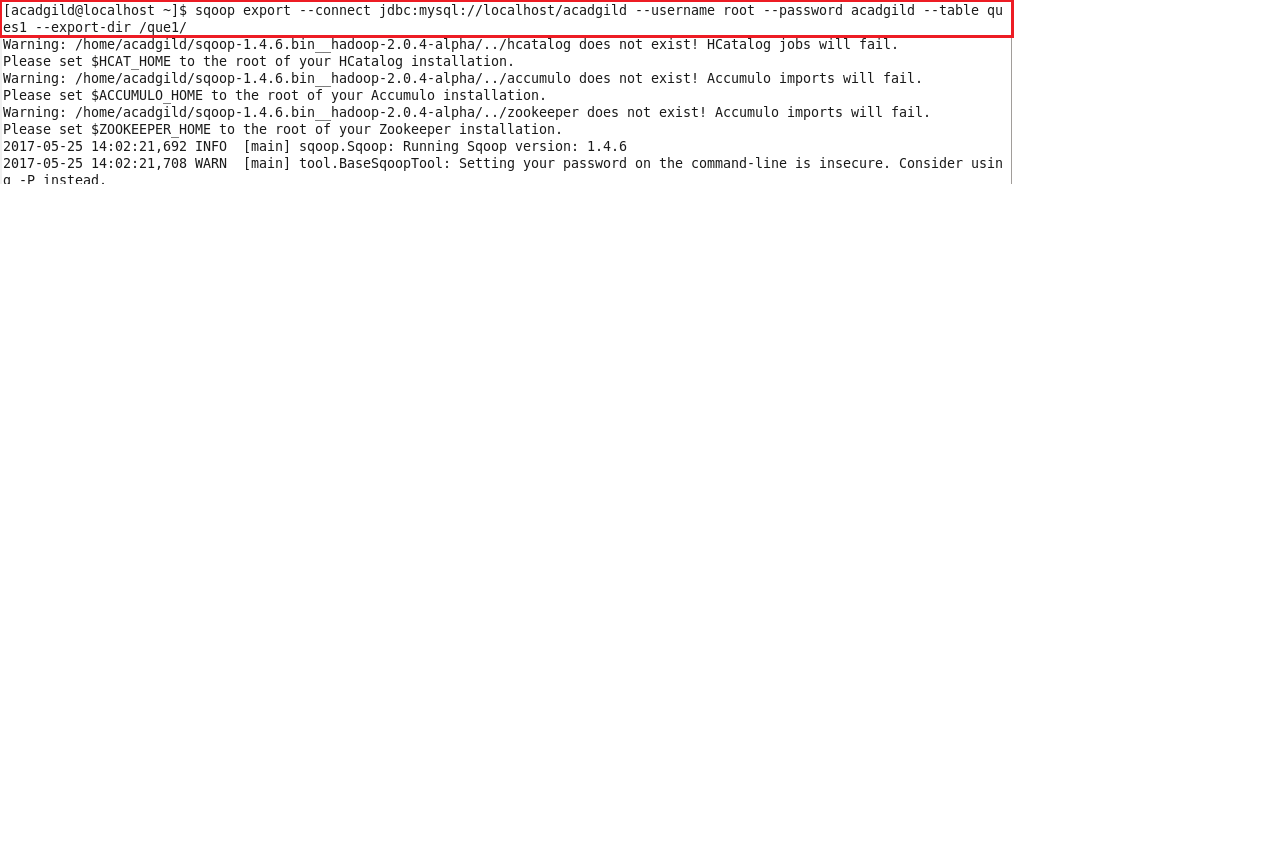


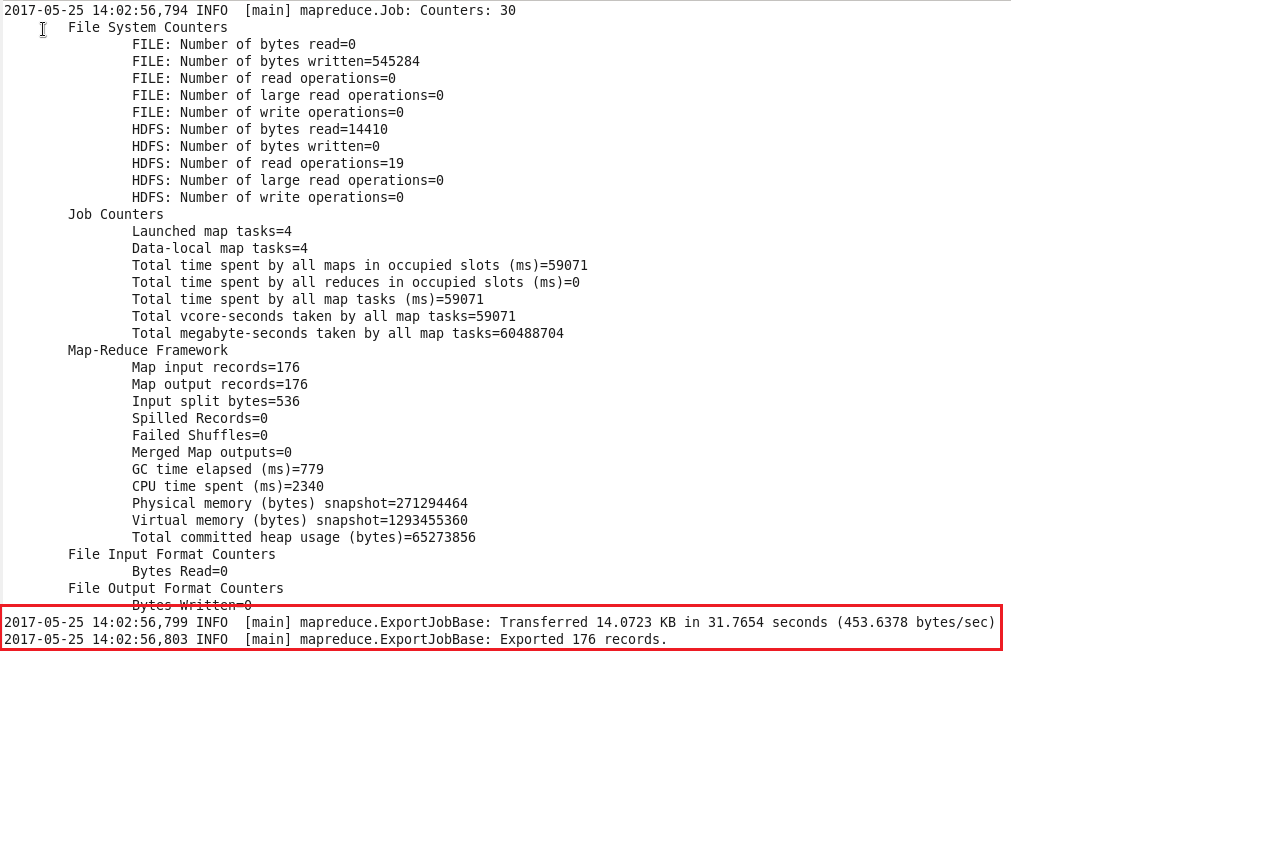
**Step 3:Exporting the results of ques1 in mysql using sqoop.**

Table ques1 is created in mysql.



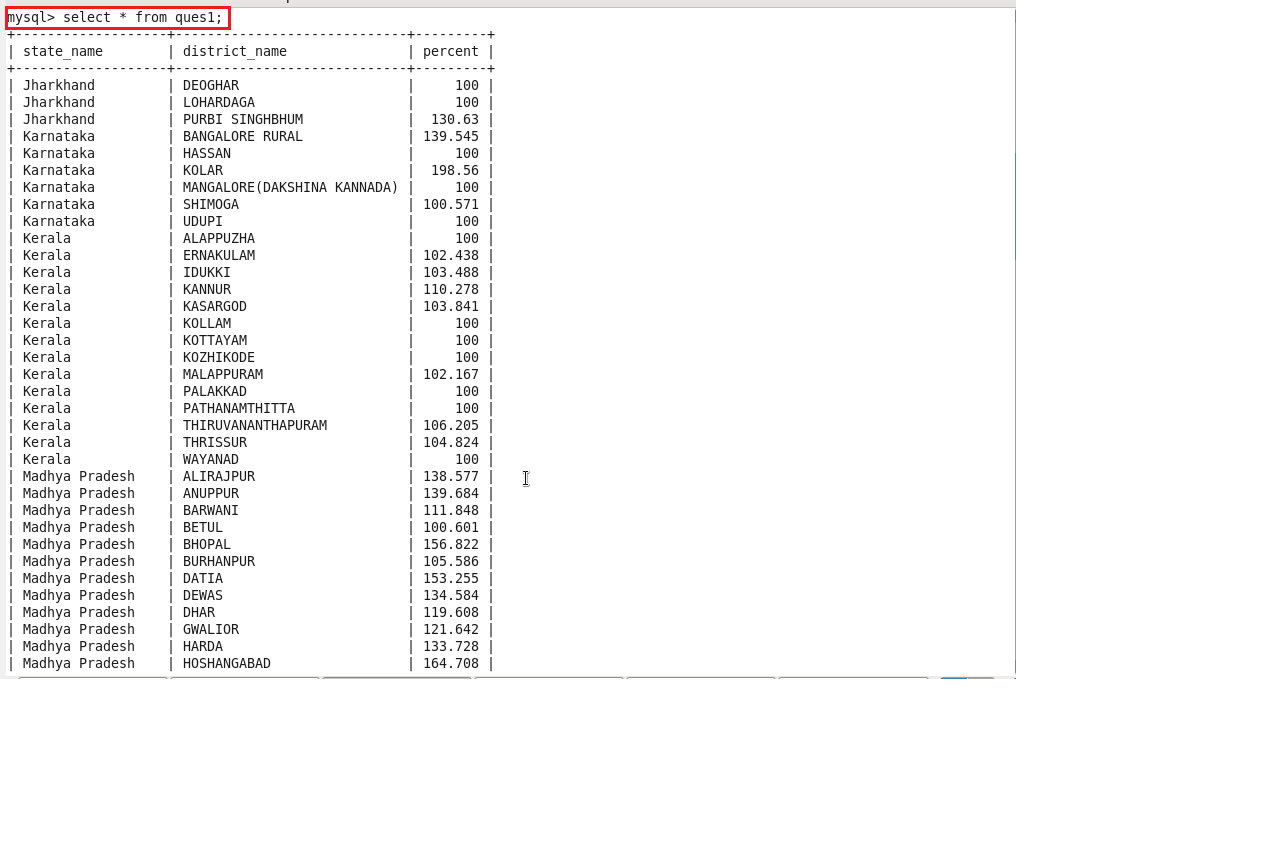
The following sqoop command is used to export 176 results records in Mysql.

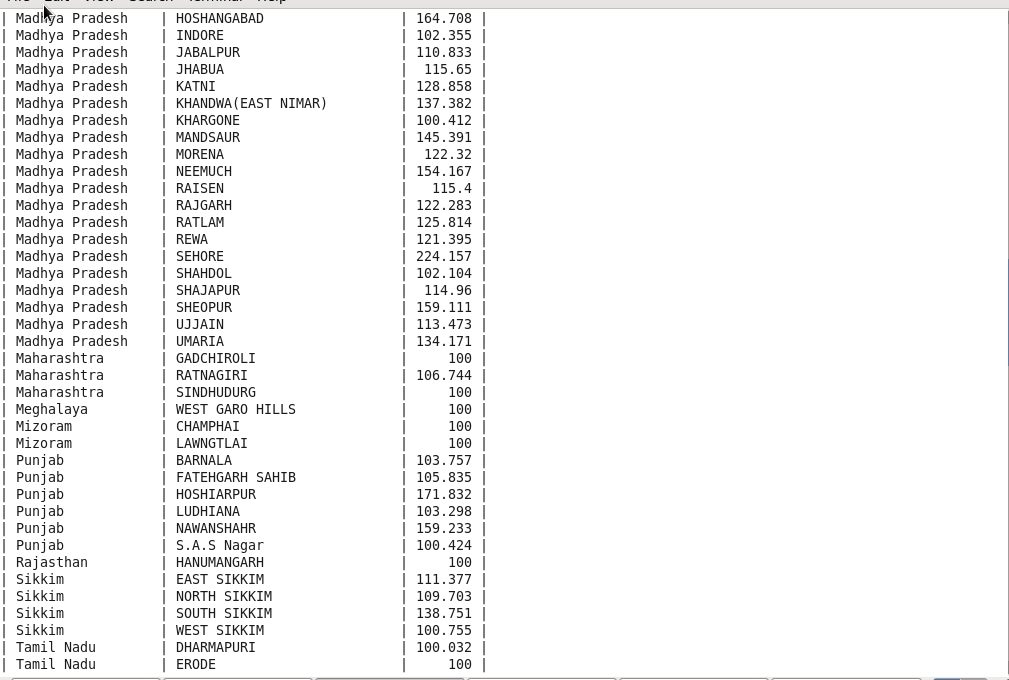


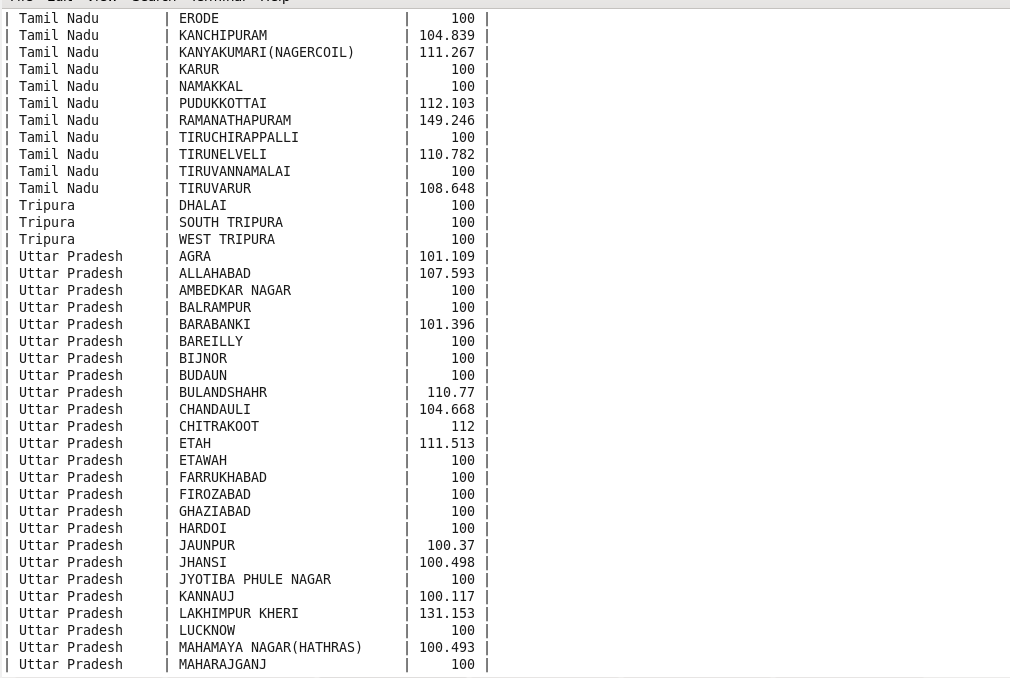


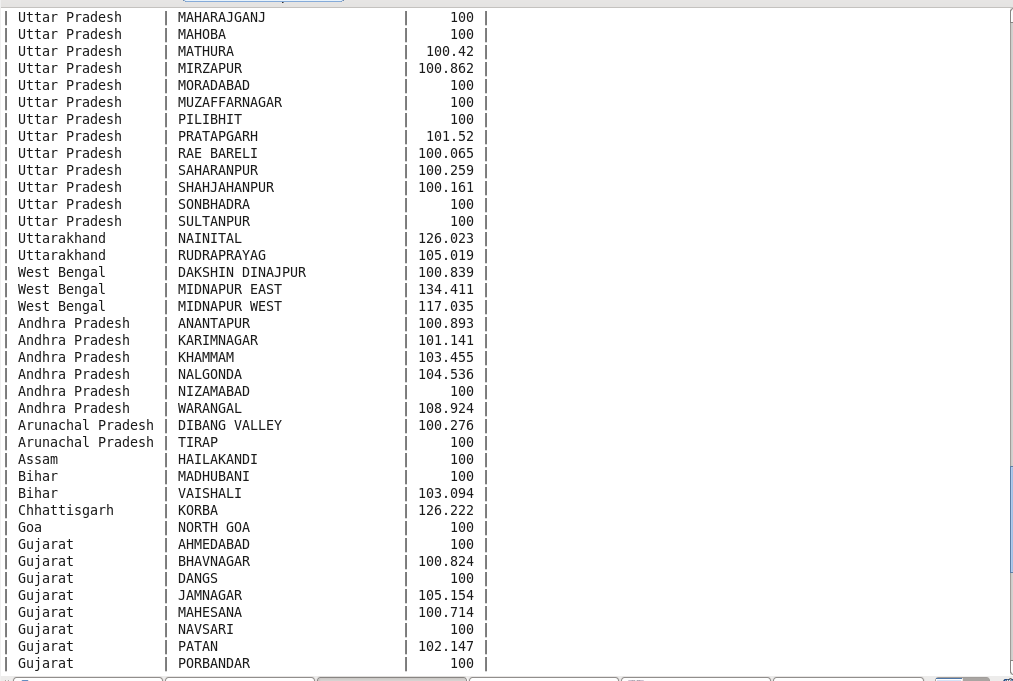
Total 176 Records exported successfully.

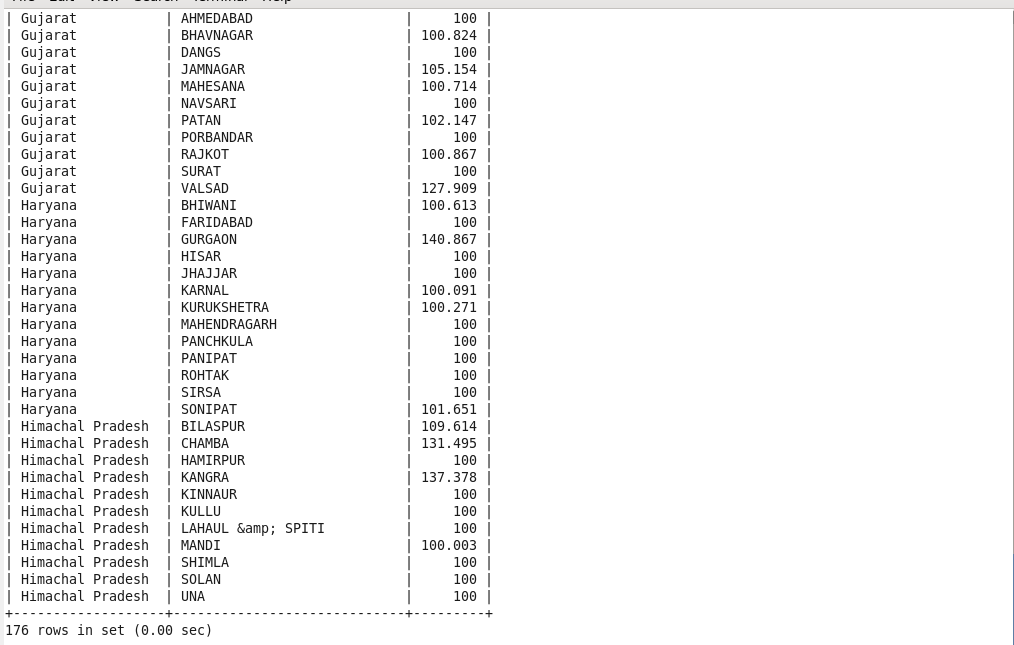
Displaying the records in Mysql table ques1.





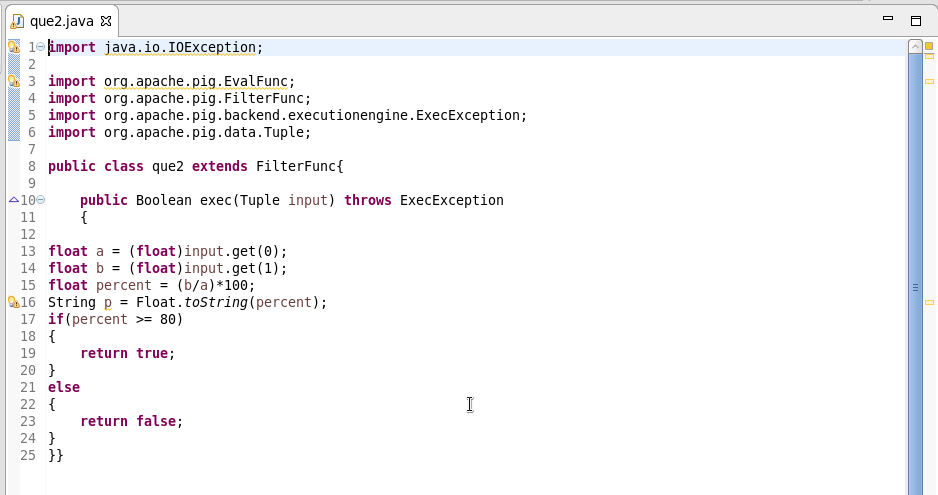






**2. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards. Export the results to MySQL using Sqoop.**

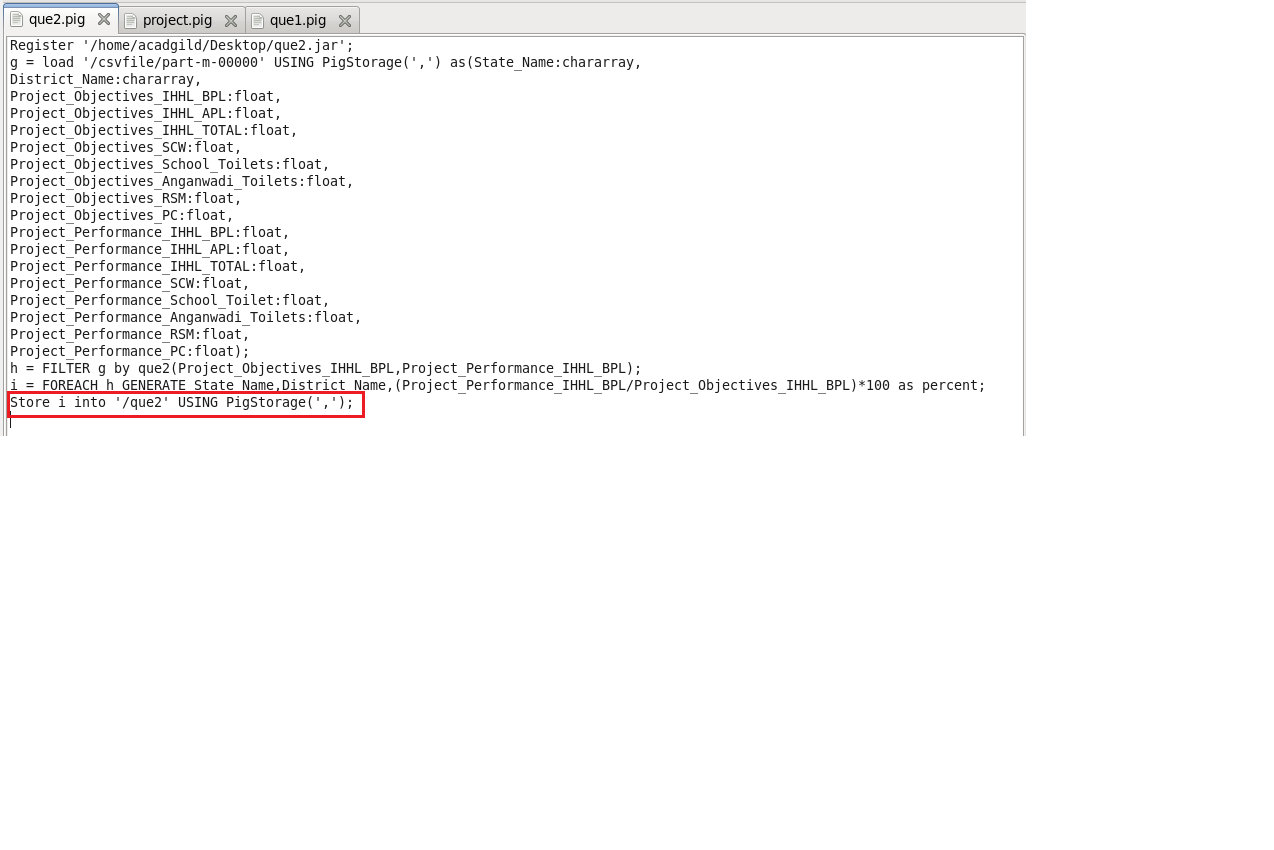
**Pig UDF**



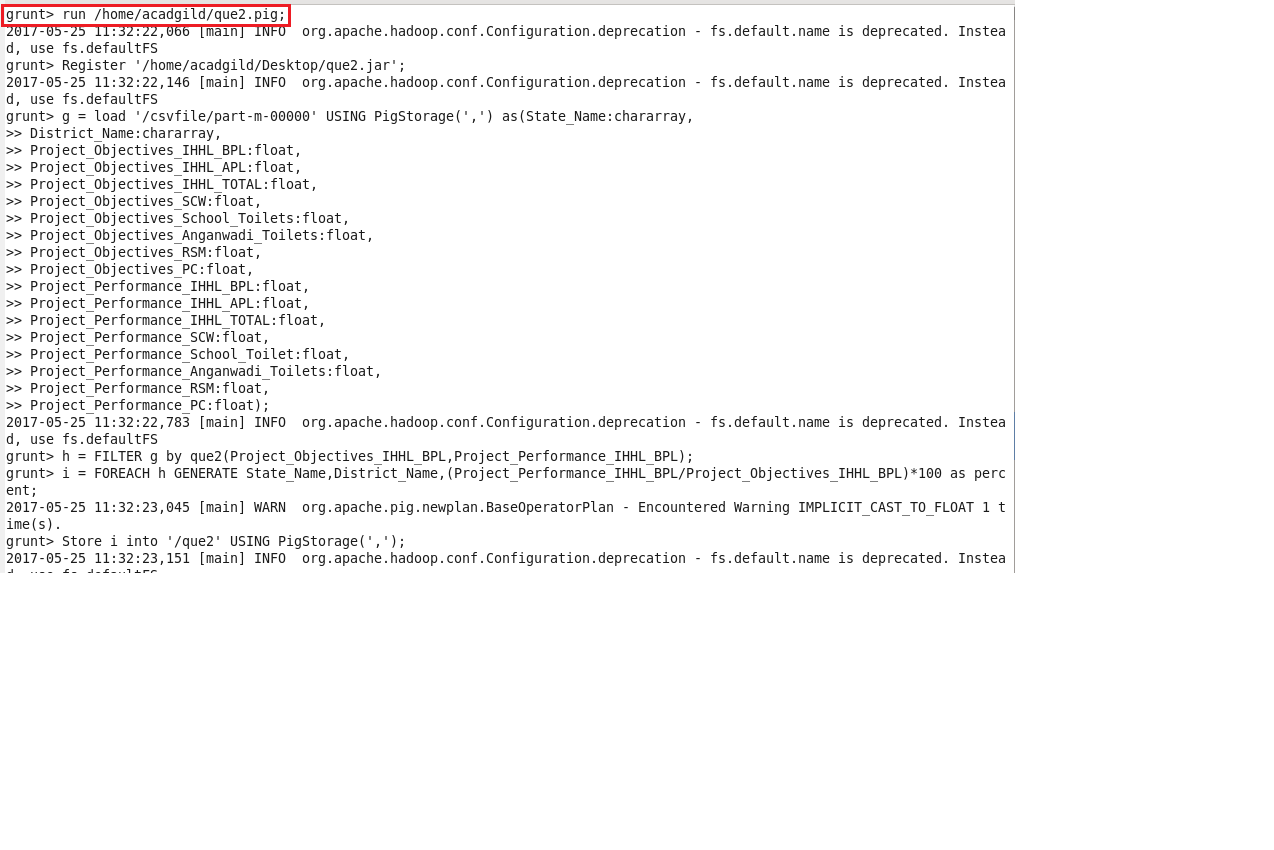
**Running Pig Script**

UDF is registered in the Pig script,dataset is loaded,filtered the records with percentage greater than 80 and generated the required columns.

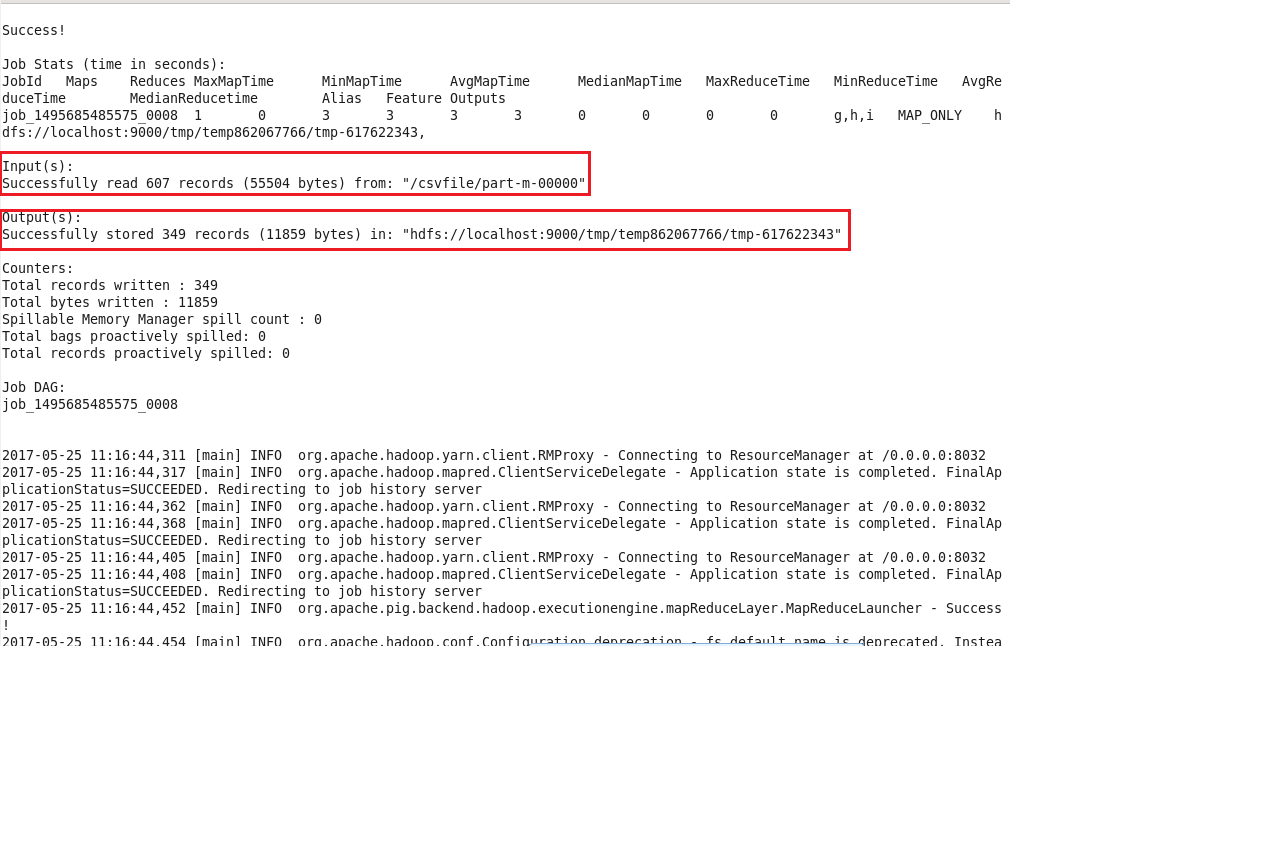
The output is stored in que2 directory.



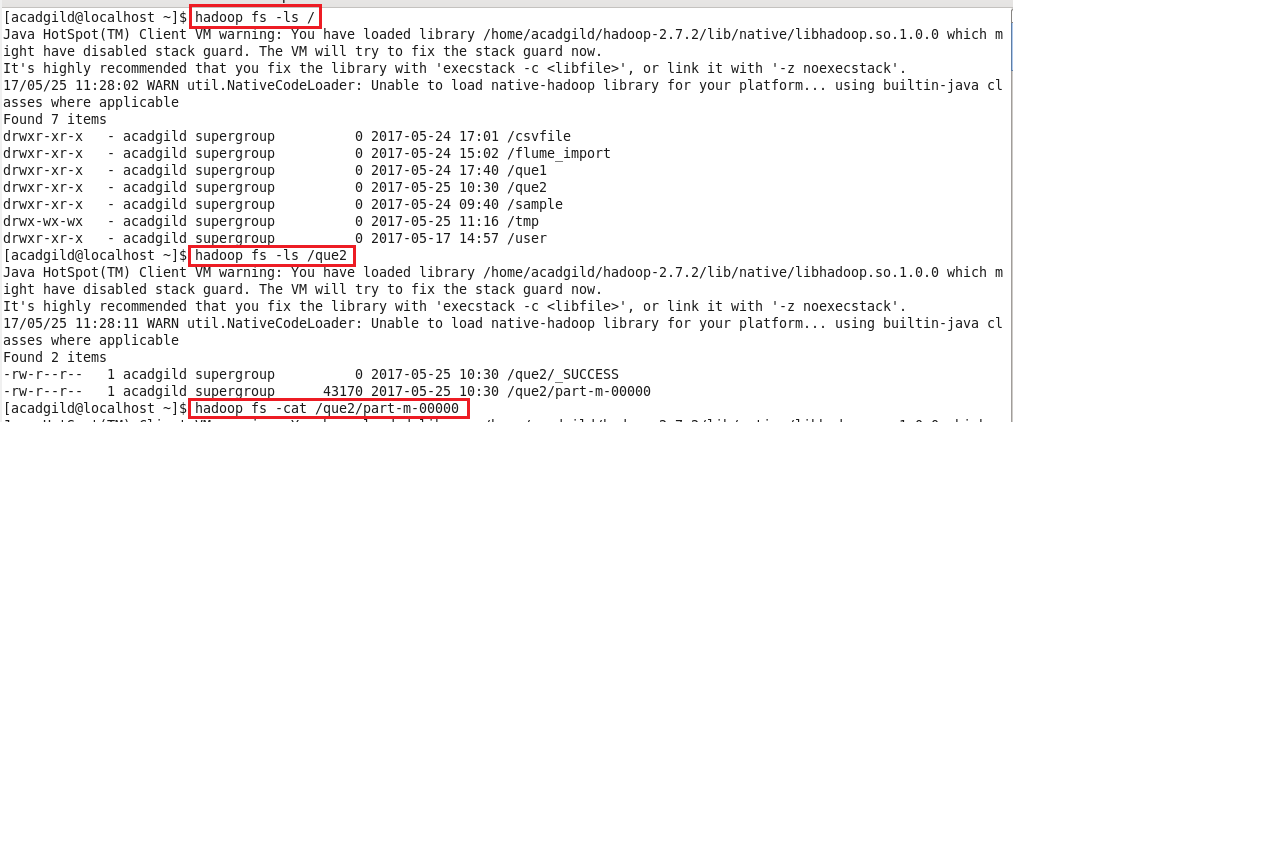
**Running Pig script**



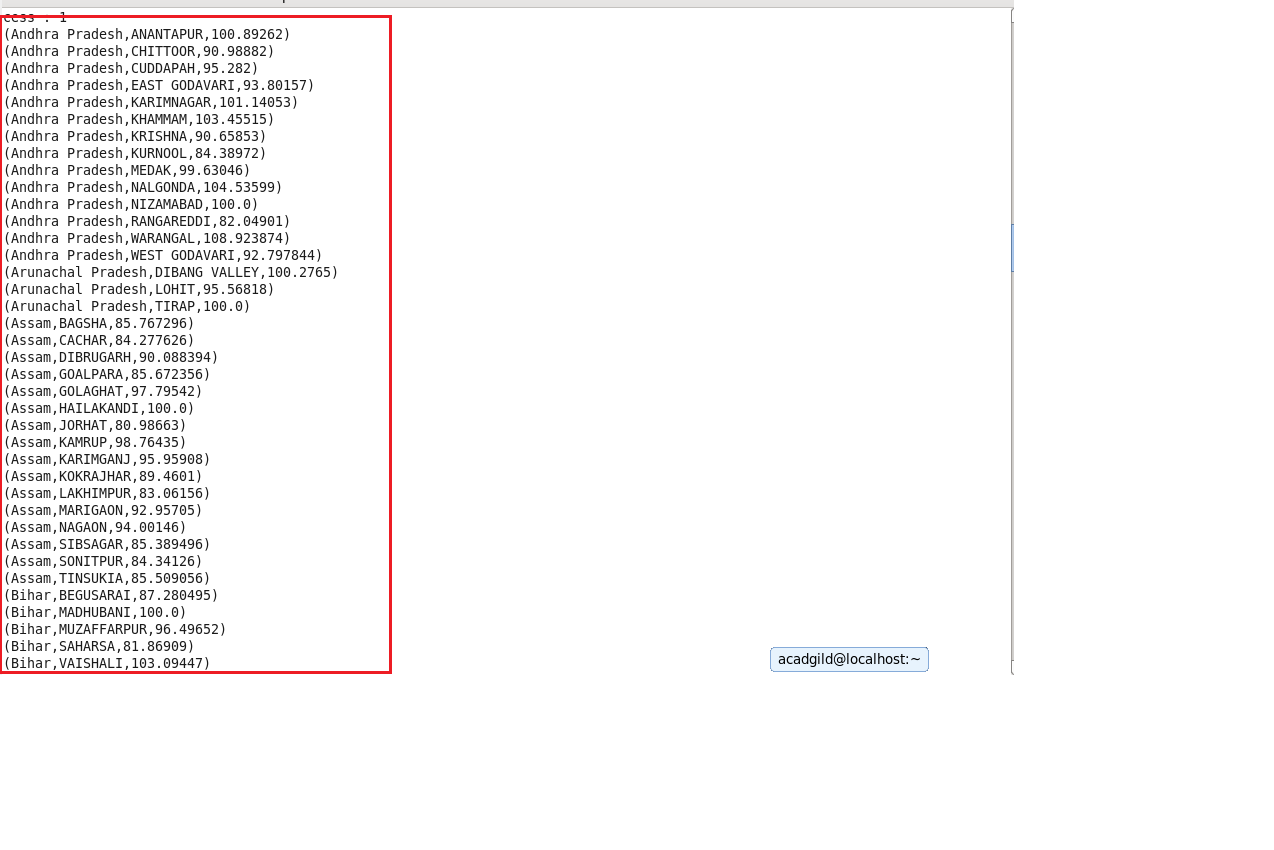
Pig script executed successfully

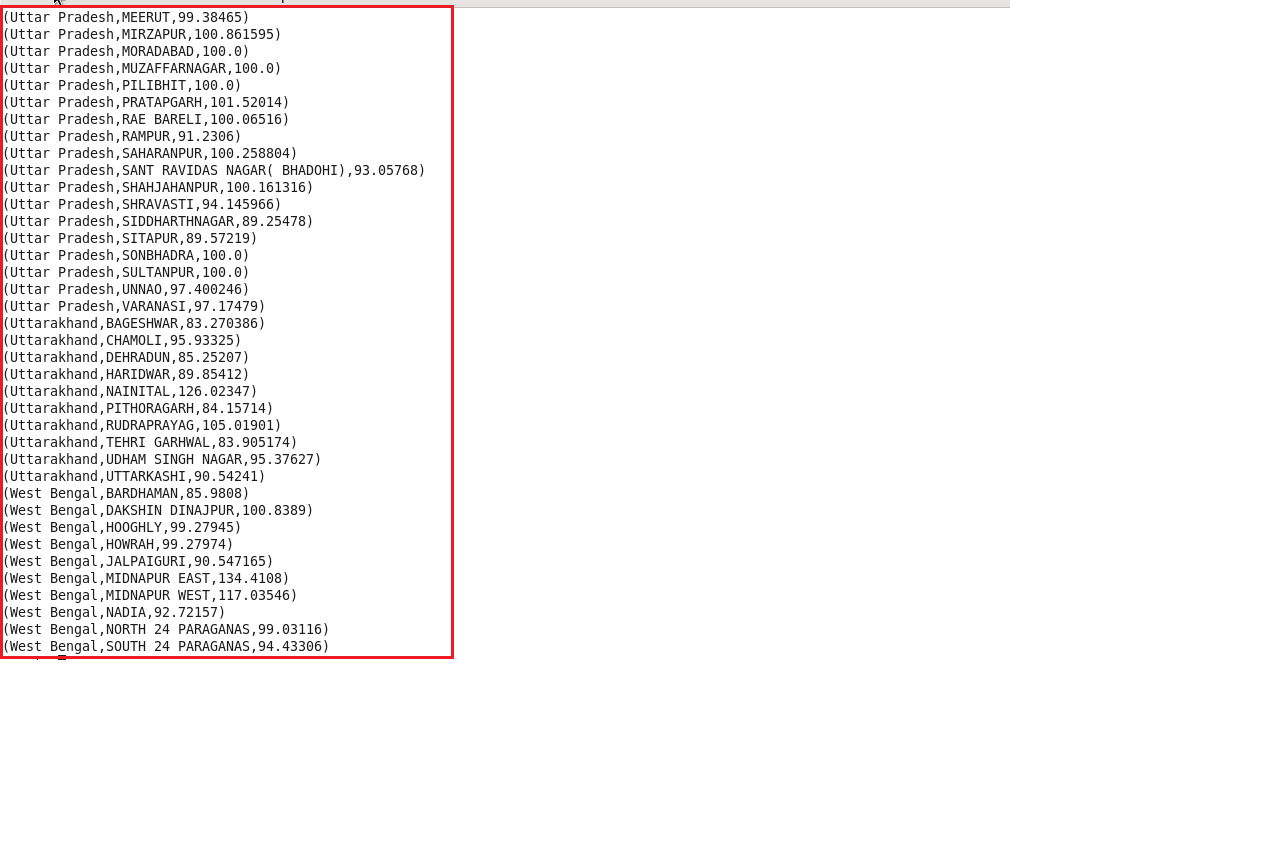


The directory que2 is created with 349 result records.



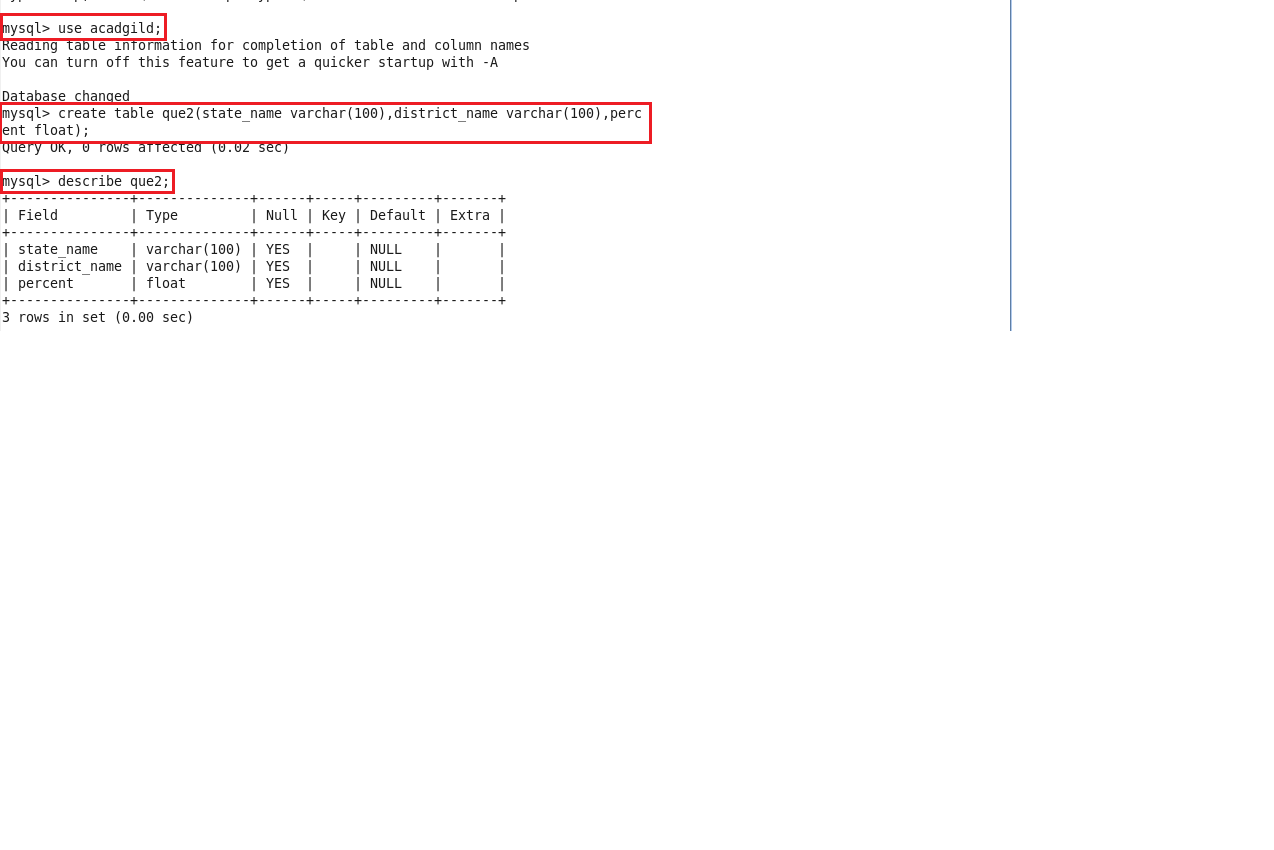
The first and last result records are displayed.



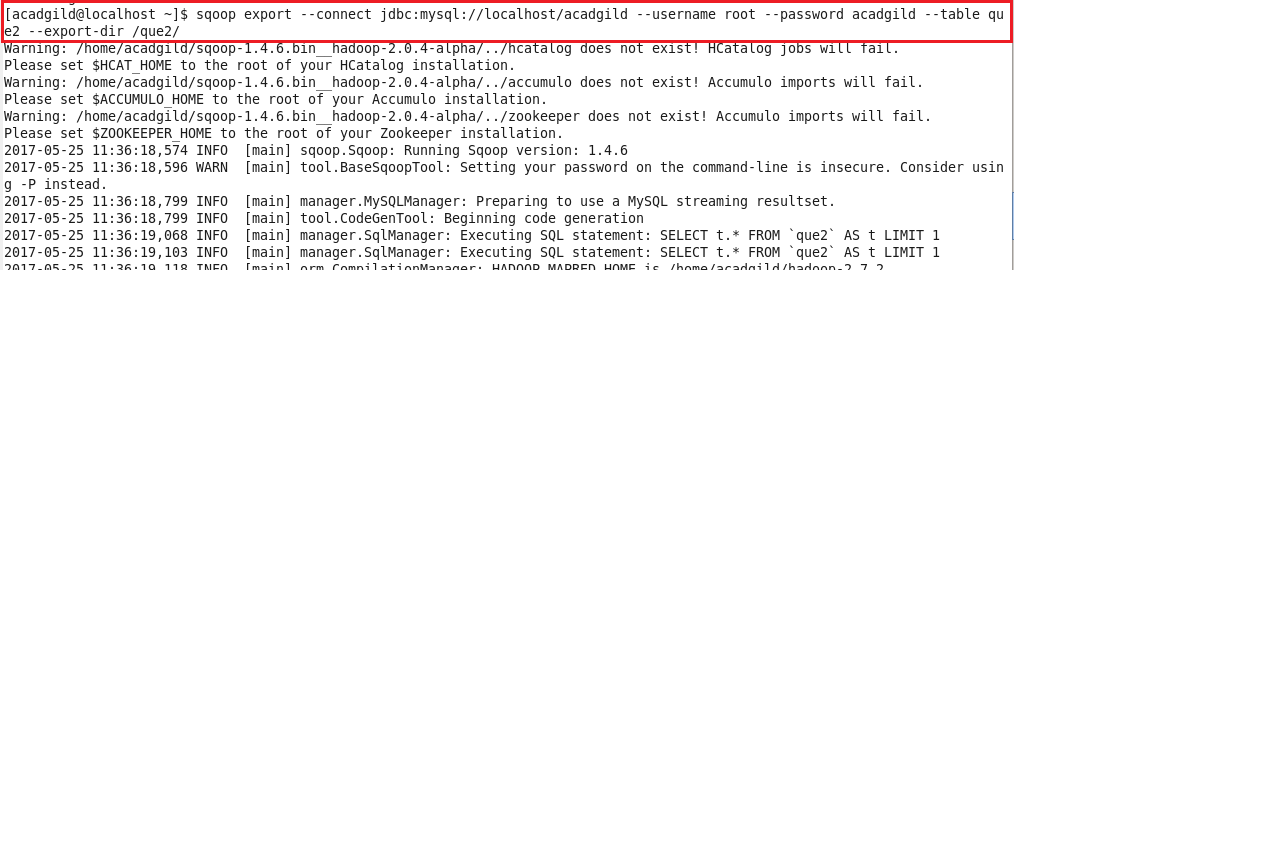


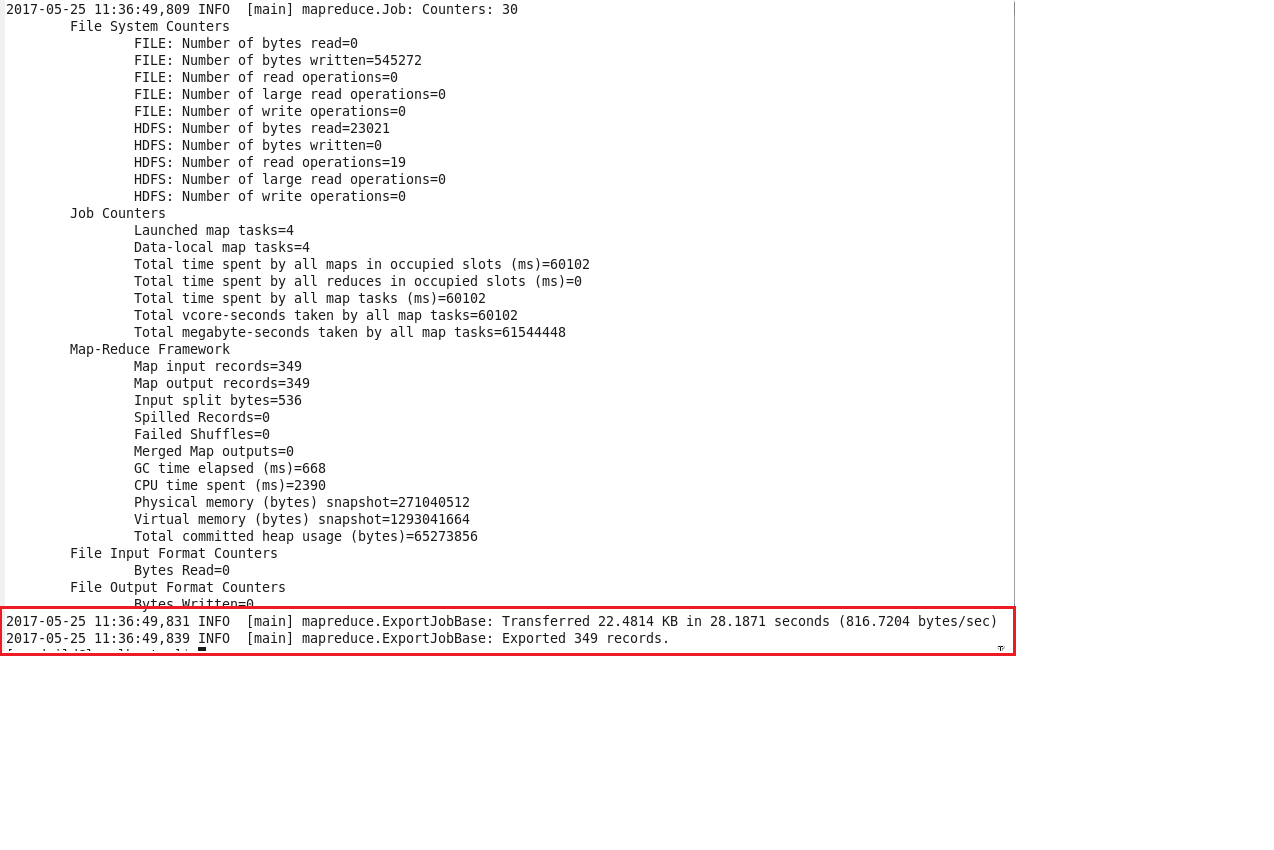
**Step 3:Exporting the results of ques2 in mysql using sqoop.**

Table que2 is created in Mysql.



The following sqoop command is used to export the result records of que2 to Mysql.





Total 349 Records exported successfully.

Displaying the records in Mysql table.

